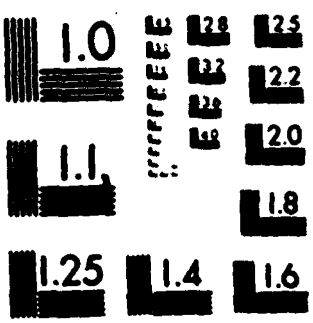
AD-A139 587 THE NORMAL APPROXIMATION AND QUEUE CONTROL FOR RESPONSE 1/1
TIMES IN A PROCES..(U) NAVAL POSTGRADUATE SCHOOL
MONTEREY CA D P GAVER ET AL. FEB 84 NPS55-84-001

UNCLASSIFIED F/G 9/2 NL

END
DATE
FILMED
4-84
DTIC


1.0
2.8
2.5
3.2
2.2
3.6
4.0
2.0
1.1
1.8
1.25
1.4
1.6

NPS55-84-001

NAVAL POSTGRADUATE SCHOOL

Monterey, California

THE NORMAL APPROXIMATION AND QUEUE CONTROL
FOR RESPONSE TIMES IN A PROCESSOR-SHARED
COMPUTER SYSTEM MODEL

by

D. P. Gaver
P. A. Jacobs
G. Latouche

February 1984

Approved for public release; distribution unlimited.

Prepared for:
Chief of Naval Research
Arlington, VA 22217

NAVAL POSTGRADUATE SCHOOL
MONTEREY, CALIFORNIA

Commodore R. H. Shumaker
Superintendent

David A. Schrady
Provost

This report was prepared with the partial support of the Probability and Statistics Program of the Office of Naval Research, Arlington, VA.

Reproduction of all or part of this report is authorized.

Donald P. Gaver

D. P. Gaver
Professor
Department of Operations Research

P. A. Jacobs

P. A. Jacobs
Associate Professor
Department of Operations Research

G. L

G. Latouche
Universite Libre de Bruxelles
Bruxelles - Belgium

Reviewed by:

Alan R. Washburn, Chairman
Department of Operations Research

Released by:

Kneale T Marshall

Kneale T. Marshall
Dean of Information and Policy
Sciences

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER NPS55-84-001	2. GOVT ACCESSION NO. A139587	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) THE NORMAL APPROXIMATION AND QUEUE CONTROL FOR RESPONSE TIMES IN A PROCESSOR-SHARED COMPUTER SYSTEM MODEL		5. TYPE OF REPORT & PERIOD COVERED Technical
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) D. P. Gaver P. A. Jacobs G. Latouche		8. CONTRACT OR GRANT NUMBER(s)
9. PERFORMING ORGANIZATION NAME AND ADDRESS Naval Postgraduate School Monterey, CA 93943		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 61153N; RR014-05-OE N0001484WR24011
11. CONTROLLING OFFICE NAME AND ADDRESS Probability and Statistics Program Office of Naval Research Arlington, VA 22217		12. REPORT DATE February 1984
		13. NUMBER OF PAGES 18
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) Unclassified
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Processor-sharing queues; job response time; central limit theorem; diffusion approximation; queue control

lambda

λ = authors

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Represent a time-shared computer system as a group of N terminals, each having submission rate λ and exponential (μ) task durations, with tasks submitted to a central (single) processor. There ~~they~~ are serviced in processor-sharing or time-sliced mode. It is shown that the R(t), the response time conditional on t , the required processing time, becomes approximately normally distributed as t increases. Similar results are derived when N increases.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

S/N 0102-LF-014-6601

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

Variations of the model consider control: an "inside," processor-shared queue services at most c tasks, others queueing first-come first-served "outside." Other possibilities are described and analyzed.

THE NORMAL APPROXIMATION AND QUEUE CONTROL FOR RESPONSE TIMES IN A PROCESSOR-SHARED COMPUTER SYSTEM MODEL

D. P. Gaver
Naval Postgraduate School

Patricia A. Jacobs
Naval Postgraduate School

Guy Latouche
University Libre Brusselles

ABSTRACT

Represent a time-shared computer system as a group of N terminals, each having submission rate λ and exponential (μ) task durations, with tasks submitted to a central (single) processor. There they are serviced in processor-sharing or time-sliced mode. It is shown that the $R(t)$, the response time conditional on t, the required processing time, becomes approximately normally distributed as t increases. Similar results are derived when N increases.

Variations of the model consider control: an "inside," processor-shared queue services at most c tasks, others queueing first-come first-served "outside." Other possibilities are described and analyzed.

1. Introduction

The abstraction of computer capacity allocation known as processor sharing is an attractive simplification of time slicing, sometimes called round-robin scheduling. The idea is well known (Kleinrock, 1976): given that j jobs or programs are at the execution stage, each receives service equal to one-jth of a time unit per time unit. In other words, if the chance that any single job, processed alone, finishes in $(t,t+h)$ is $\mu h + o(h)$, (exponential-Markov service), then the chance that a particular ("tagged") job in the company of $(j-1)$ others finishes in $(t,t+h)$ is $\mu(h/j) + o(h)$ as $h \to 0$. Processor sharing of the above type tends to be equitable in that it permits short jobs access to processing even if they arrive after, and queue with, longer jobs.

Apparently the first study of delays to arriving and queueing jobs under processor sharing was conducted by Coffman, Muntz, and Trotter (1970). They assumed a steady state M/M/1 system with processor sharing, and were able to determine properties of the response time, R, given the processing time required by the arriving job. Other papers have also appeared.

A recent paper by D. Mitra (1981) analyzes response time, R, under the assumption of a closed system. Idealize the behavior of a system of N terminals and a single computer as a classical machine-repair situation: each thinking terminal (failure-prone machine) applies for computer service at rate λ, and queued or waiting jobs are served at rate μ as long as any jobs are present. Markov assumptions are made throughout, so $X(t)$, the number of jobs at the service stage, is a birth and death process with transition rates

$$\begin{aligned} X(t) = j \rightarrow X(t+h) &= j + 1 : \quad \lambda_j h + o(h) \qquad (1.1) \\ \rightarrow X(t+h) &= j - 1 : \quad \mu_j h + o(h) \\ \rightarrow X(t+h) &= j \qquad : \quad 1 - (\lambda_j+\mu_j)h + o(h) \end{aligned}$$

and in particular $\lambda_j = \lambda(N-j)$, $\mu_j = \mu$ for $j \geq 1$, otherwise being zero. Let processor sharing govern service effort allocation. In Mitra (1981) the distribution of response time is characterized, and the moments (e.g. mean and variance) are found under interesting conditions, such as that the tagged job requiring t time units of processing arrives to find $j - 1$ accompanying jobs; the conditional response time, given only processing requirement t, is given particular attention.

In this paper the previous analysis is generalized and extended. We introduce the idea of processor sharing in an arbitrary birth and death process environment, thus allowing quite general terminal-computer interactions to be represented. In the process, the meaning of "system state at the moment of tagged job arrival" is clarified; see also recent work of Lavenberg and Reiser (1981). Response time characteristics are computed under the assumption that processor-sharing service rates are processor-state-dependent in a more general way than that described earlier; this allows for approximate representation of overhead penalties and also of job scheduling. Other characteristics of tagged job response are also studied, e.g. the accumulated processing work, $W(\tau)$, actually performed on that job by elapsed time $\tau(\tau < t$ = required processing time) following job introduction; note that $W(R) = t$, so the first passage of $W(\tau)$ to t is actually the response time.

Although differential equations may be obtained for transforms of $W(\tau)$ under various initial conditions, and hence, implicitly, for its distribution, the results are far from being explicit and informative. However, central limit theorems for additive functionals of Markov processes, or for cumulative processes, allow the conclusion that the accumulated work accomplished by fixed time τ on a "long" job is approximately normally distributed (Gaussian). This fact in turn allows the conclusion that the response time for a "long" job is also approximately normally distributed. Additionally, the normal approximation may be shown to be valid for our simple model--and probably for others as well--when the number of competing terminals becomes large, i.e. under heavy traffic conditions. The quality of the normal approximations for finite job lengths is currently being assessed by simulation methods.

In the latter part of this paper we describe queue control in a processor-sharing environment. The expedient is to limit the total number of jobs allowed simultaneous processor-shared service at an "inside" queue, with any excess in "first-come, first-served" status in an "outside" queue. Long jobs are also shifted from inside to outside by a sampling mechanism. It is shown that long jobs are favored by a small inside span, c (c being the number simultaneously processor-shared), while short jobs are favored by large c.

2. Mean Response Times

Begin by describing differential equations for the mean response time to be experienced by a tagged, particular, arriving job. Other moments satisfy very similar equations.

(a) Conditioning on Required Time and System State.

Throughout what follows Markovian assumptions are made: service times at the computer are independent and exponential (μ). Generalizations to phase-type distributions are apparently possible.

Let R refer to the response time of a newly arrived job, and

$$m_j(t) = E[R \mid X(0) = j,\ W(R) = t] \ , \tag{2.1}$$

the conditional expectation of response time, given that the job is initially in the company of j others (arrives to find $j - 1$ present) and requires "work" or processing time equal to t . Let $\lambda_j h$ (resp. $\mu_j h$) for small h identify the infinitesimal generator of the accompanying process, so transition rates are as in (1.1).

Consider the possible system changes in $(0,h)$, and subsequently; the following results occur: $\left\{\text{let } \tilde{\mu}_j = \mu(j-1)\frac{r(j)}{j}\right\}$:

$$\begin{aligned} m_j(t) = h + m_j\left(t - \frac{r(j)}{j}\cdot h\right)[1 - (\lambda_j + \tilde{\mu}_j)h] \\ + \lambda_j h m_{j+1}\left(t - \frac{r(j)}{j}\cdot h\right) + \tilde{\mu}_j h m_{j-1}\left(t - \frac{r(j)}{j}\cdot h\right) + o(h) \ . \end{aligned} \tag{2.2}$$

The term $r(j)$ represents the fraction of time the processor actually spends processing when there are j jobs being processed.

Allowing $h \to 0$ one finds the differential equations

$$\frac{r(j)}{j} m_j'(t) = 1 - (\lambda_j + \tilde{\mu}_j)m_j(t) + \lambda_j m_{j+1}(t) + \tilde{\mu}_j m_{j-1}(t) \ . \tag{2.3}$$

This is a standard system of linear differential equations with constant coefficients; initial conditions are $m_j(0) = 0$ for all j .

(b) Conditioning on Required Time.

If one removes the condition that $X(0) = j$ in accordance with the stationary distribution appropriate for an arriving job it follows that the expected response time is linear in the required processing time, t . This holds for quite general birth-and-death process models, and not just for the simple machine-repair setup; see Cohen [1979]. Here is the derivation, in outline.

First, observe that the long-run distribution of $X(0)$, the number of jobs present just after the tagged job enters, is

$$q_j = c\pi_{j-1}\lambda_{j-1} = c\pi_j\mu r(j) \qquad j = 1,2,\ldots,N \,, \tag{2.4}$$

where c is selected so that the q_j's sum to one. Recall that $\pi_j = \pi_0 \frac{\lambda_0\lambda_1\cdots\lambda_{j-1}}{\mu_1\mu_2\cdots\mu_j}$ is the stationary distribution (assumed to exist) of the Markov chain $X(t)$ defined by (1.1) with $\mu_j = \mu r(j)$. This is intuitively apparent, but a formal proof can be based either upon an embedded Markov chain formulation, or upon the theory of additive functionals of a Markov process; see Çinlar ((1975), pp. 269-271). The distribution $\{q_j\}$ has also been given by Kelly, (1979), p. 12.

Use (2.4) to remove the condition that $X(0) = j$; put

$$m(t) = E_{X(0)}E[R|X(0),\ W(R) = t] = \sum_{j=1}^{N} q_j m_j(t) \,. \tag{2.5}$$

Then in terms of the differential equations (2.3); after multiplying through by $j/r(j)$ one obtains $(m_0(t) \equiv 0)$

$$m'(t) = \sum_{j=1}^{N} \frac{1}{r(j)}\, q_j + \sum_{j=1}^{N} q_j \frac{1}{jr(j)} [-(\lambda_j + \tilde{\mu}_j)m_j(t) + \lambda_j m_{j+1}(t) + \tilde{\mu}_j m_{j-1}(t)]$$

$$= \sum_{j=1}^{N} \frac{1}{r(j)}\, q_j \,. \tag{2.6}$$

Thus it follows that the long-run conditional expected response time is linear in the processing time requirement:

$$E[R|W(R) = t] = t\sum_{j=1}^{N} \frac{1}{r(j)}\, q_j = tE\left[\frac{X(0)}{r(X(0))}\right] \,. \tag{2.7}$$

Apparently no such simple form exists for $\mathrm{Var}[R|W(R) = t]$, although Mitra (1981) has given a formula for a particular case. It will be seen, however, that the above variance is indeed proportional to t if t is large.

3. Total Work Completed on a Tagged Job in a Fixed Time

Turn attention now to $W(t)$, the total work expended by the computer on the tagged job by time t after its arrival, given that the tagged job requires exactly t time-units of work for completion. If when the job arrives there are $X(0) = j$ customers present, then

$$W(\tau) = \int_0^{\tau} \frac{r(X(u))}{X(u)}\, du, \qquad X(0) = j \geq 1 . \tag{3.1}$$

(a) The Laplace transform of $W(\tau)$.

Here is the derivation of a differential equation for the Laplace transform of $W(\tau)$:

$$\phi_j(s,\tau;t) = E[e^{-sW(\tau)} | X(0) = j,\ R = t] , \qquad \text{for} \quad j \geq 1 \tag{3.2}$$

Now argue in a manner analogous to the discussion prior to (2.2) to write a backward equation: for $0 < h < t$

$$\begin{aligned} \phi_j(s,\tau;t) = {} & (e^{-sr(j)h/j})[1-(\lambda_j+\tilde{\mu}_j)h]\phi_j(s,\tau-h;t) \\ & + (e^{-sr(j)h/j})\lambda_j h\phi_{j+1}(s,\tau-h;t) \\ & + (e^{-sr(j)h/j})\tilde{\mu}_j h\phi_{j-1}(s,\tau-h;t) + o(h) , \end{aligned} \tag{3.3}$$

leading to

$$\frac{d\phi_j}{d\tau} = -(\lambda_j + \tilde{\mu}_j + sr(j)/j)\phi_j + \lambda_j\phi_{j+1} + \tilde{\mu}_j\phi_{j-1} . \tag{3.4}$$

Initial conditions are

$$\phi_j(s,0;t) = E[e^{-sW(0)} | X(0) = j,\ R = t] = 1 , \qquad t > 0 , \tag{3.5}$$

since initially $W(0) = 0$, regardless of the job requirements or the initial environment.

(b) A central limit theorem for $W(\tau)$.

Examination of (3.1) shows that $W(\tau)$ involves sums of contributions to work accumulated while the system inhabits various states during the period $(0,\tau)$. This suggests that, at least for "long" jobs, i.e. such that required processing time $t \to \infty$, one can anticipate a nearly-Normal distribution for $W(\tau)$. An appropriate central limit theorem that establishes this for finite birth-and-death models can be found in Keilson ((1979), p. 121); call this Theorem K. Alternatively, one can make use of the theory of cumulative processes, see Cox ((1962), pp. 99-101); the latter development is adaptable to models more general than the simple birth-and-death process.

In order to apply Theorem K redefine the infinitesimal generator (1.1) to describe the behavior of the accompaniment, $X'(t)$, of the tagged customer; note that the relevant generator is now

$$X'(t) = j \to X'(t+h) = j+1 : \quad \lambda_j' h + o(h)$$

$$\equiv \lambda_{j+1} h + o(h) \tag{3.6}$$

$$\to X'(t+h) = j-1 : \quad \mu_j' h + o(h)$$

$$\equiv \mu_{j+1}\left(\frac{j}{j+1}\right) h + o(h) \tag{3.7}$$

$$\to X'(t+h) = j \quad : \quad 1-(\lambda_j' + \mu_j')h + o(h) \tag{3.8}$$

for $j = 0,1,2,\ldots, N' = N-1$. Then

$$W(t) = \int_0^t f(X'(u))du \equiv \int_0^t \frac{f(X'(u)+1)du}{X'(u)+1} \tag{3.9}$$

and theorem K states that

$$\frac{W(t) - \zeta t}{\sigma\sqrt{t}} \xrightarrow{(D)} N(0,1) \; ; \tag{3.10}$$

the constants ζ and σ^2 are such that

$$\zeta = \sum_{j=0}^{N'} f(j)\pi_j' \tag{3.11}$$

$$\sigma^2 = 2[f(0), f(1), \ldots, f(N')] \begin{bmatrix} \pi_0' & & 0 \\ & \pi_1' & \\ 0 & & \pi_{N'}' \end{bmatrix} \mathbf{Z} \begin{bmatrix} f(0) \\ f(1) \\ \vdots \\ f(N') \end{bmatrix} , \tag{3.12}$$

where in the present case the definition of $f(\cdot)$ is implicit in (3.9), and $\mathbf{Z}$ is the matrix

$$\mathbf{Z} = \frac{1}{\gamma}\{[\mathbf{I}-\mathbf{A}+\mathbf{L}]^{-1} - \mathbf{L}\} \tag{3.13}$$

$\mathbf{I}$ being the identity, and

$$\mathbf{\Pi} = \begin{vmatrix} \pi_0' & \pi_1' \cdots \pi_{N'}' \\ \pi_0' & \pi_1' \cdots \pi_{N'}' \\ \pi_0' & \pi_1' \cdots \pi_{N'}' \end{vmatrix} , \tag{3.14}$$

$\mathbf{\Pi}$-rows are steady-state probabilities for the accompaniment, and $\mathbf{A}$ is defined as follows:

$$A_{0,1} = \lambda_0'/\gamma , \quad A_{0,0} = 1 - \lambda_0'/\gamma$$

$$A_{j,j+1} = \lambda_j'/\gamma , \quad A_{j,j-1} = \mu_j'/\gamma , \quad A_{j,j} = 1 - \nu_j/\gamma \tag{3.15}$$

$$\nu_j = \lambda_j' + \mu_j' ; \quad \gamma = \max_j \nu_j .$$

$$(\nu_0 = \lambda_0', \nu_{N'} = \mu_{N'}')$$

(c) A central limit theorem for response time, $R(t)$.

A graph of $W(\tau)$ vs. τ starts with $W(0) = 0$ and increases in random straight-line segments until $W(\tau) = t$. The value of τ at which this occurs, $\tau(t) \equiv R(t)$, is the first-passage time to t of the work process $\{W(\tau), \tau\geq 0\}$, and is the required response time, so

$$P\{W(\tau)<t\} = P\{R(t)>\tau\} . \tag{3.16}$$

Now invoke the previous theorem (3.10) concerning asymptotic normality of $W(\tau)$ and a standard argument of renewal theory, cf. Karlin and Taylor ((1979), pp. 208-209) to see that if $t = \xi\tau + \sqrt{\sigma^2\tau}x$, then, as $t \to \infty$, $\tau \sim \frac{t}{\xi} - \sqrt{\frac{\sigma^2 t}{\xi^3}}\, x$, from which it follows that

$$\frac{R(t) - \alpha t}{\sqrt{\beta^2 t}} \equiv \frac{R(t) - t/\xi}{\sqrt{\frac{\sigma^2}{\xi^3} t}} \overset{(D)}{\to} N(0,1) . \tag{3.17}$$

4. Heavy Traffic Analysis of the Response Time of a Processor-Shared Job

This section investigates the problem of delay of a tagged job requiring t units of processing time when it is accompanied by many others, i.e. is in a heavily loaded system. Restrict attention to the machine repair model in which $\lambda_j = \lambda(N-j)$ and $\mu_j = \mu$, and omit the effect of $r(j)$, i.e. $r(j) \equiv 1$. Let there be N terminals and one processor, with λ^{-1} being the expected terminal think time (exponentially distributed), $\mu = N\mu'$ being the processing rate of arriving jobs; λ and μ' are fixed but N is large and the service rate scaling by N is required in order that queue size be of order N.

Now utilize the fact (Iglehart (1965), and Gaver and Lehoczky (1976)) that if $X(t)$ is the number of jobs at the processing stage at t then $X(t)$ can be approximated by a diffusion process:

$$X(t) = N\,a(t) + \sqrt{N}\,Y(t) \tag{4.1}$$

where $a(t)$ is a deterministic function of time and $\{Y(t)\}$ is, for the present model, a particular Ornstein-Uhlenbeck process. It turns out that when $N \to \infty$

$$\frac{da(t)}{dt} = \lambda(1-a(t)) - \mu' \tag{4.2}$$

or

$$a(\infty) = 1 - \frac{\mu'}{\lambda} \tag{4.3}$$

which is feasible if $\lambda > \mu'$, i.e. under heavy traffic conditions. Furthermore

$$dY(t) = -\lambda Y(t)dt + \sqrt{\lambda(1-a(t)) + \mu'}\,dB(t)\,, \tag{4.4}$$

$\{B(t), t \geq 0\}$ being the standard Wiener process. In the long run,

$$dY(t) = -\lambda Y(t)dt + \sqrt{2\mu'}\,dB(t)\,. \tag{4.5}$$

It is in the environment $X(t)$ described by (4.1) that the tagged job enters. It encounters competition for processor-shared service, and so its accumulated work completed by fixed time τ is essentially

$$W(\tau) = \int_0^\tau \frac{du}{X(u)}\,. \tag{4.6}$$

Apply the approximation and expand to second order terms in N, the number of terminals, to find

$$W(\tau) \simeq \int_0^\tau \frac{du}{Na(u)\left[1 + \frac{Y(u)}{a(u)\sqrt{N}}\right]} \simeq \int_0^\tau \frac{du}{Na(u)} - \int_0^\tau \frac{\sqrt{N}\, Y(u)du}{(Na(u))^2} \tag{4.7}$$

For simplicity, and to enable comparisons with previous results, let $a(v) = a(\infty)$, so the tagged job arrives in the steady state. Expression (4.6) then says that for the approximation advanced here,

$$W(\tau) = \int_0^\tau \frac{du}{E[X(\infty)]} - \frac{\sqrt{N}}{(E[X(\infty)])^2} \int_0^\tau Y(u)du\,, \quad 0 \leq \tau \leq t \tag{4.8}$$

so the expected amount of work done on the tagged job is nearly $\tau/E[X(\infty)]$, and the actual distribution of total work done is approximately Gaussian (integral of an Ornstein-Uhlenbeck process), where the Gaussian property results from the assumption of many accompanying jobs, and not necessarily because the tagged job is long.

Standard calculations applied to (4.7) show that, as $\tau \to \infty$, $E[\int_0^\tau Y(u)du] \simeq 0$ and $Var[\int_0^\tau Y(u)du] \simeq (2\mu'/\lambda^2)\tau$, so the normal approximation to accumulated work $W(\tau)$ has the parameters

$$\xi = \frac{1}{[Na(\infty)]}\,, \qquad \sigma^2 = \frac{2\mu'}{\lambda^2} \cdot \frac{N}{[Na(\infty)]^4} = \frac{2\mu}{\lambda^2[Na(\infty)]^4} \tag{4.9}$$

from which it follows that the parameters of the normal approximation to $R(t)$ are

$$\alpha = \frac{1}{\xi} = [Na(\infty)] \simeq E[X(\infty)]$$

$$\beta^2 = \frac{2\mu}{\lambda^2}\frac{1}{Na(\infty)} \simeq \frac{2\mu}{\lambda^2}\frac{1}{E[X(\infty)]} = \frac{2}{\lambda}\left(\frac{N - E[X(\infty)]}{E[X(\infty)]}\right) . \tag{4.10}$$

These formulas state that if think (demand) rate λ is very large then, since $E[X(\infty)] \to N$, the variance of response time diminishes, while of course expected response time increases like N . This is plausible since in extremely heavy traffic all terminals compete, and the tagged job gets a steady $(1/N)^{\underline{th}}$ of a quantum. For smaller λ the expected response time drops with $E[X(\infty)]$, but response time variance increases.

The above derivations are informative but not rigorous. Semigroup methods of Berman (1979) can be applied to place the results on a mathematically solid basis. Numerical assessment of the results is also of interest.

5. Numerical Comparisons

In this section a brief investigation is reported of the numerical agreement between the very simple formulas from heavy traffic theory for the parameters of the accumulated work distribution and those of direct Markov-chain cumulative process theory origin.

Parameters of Total Work

Examples. Let $\mu = 1$, $N = 25, 50$, with λ varying.

$N = 25$

Rates(λ):	0.01	0.0222...	0.05	0.10	0.15
ξ(M.Ch.)	0.77	0.51	0.16	0.067	.055
(Diffus.)	—	—	0.20	0.067	0.055
σ^2(M.Ch.)	0.44	0.75	0.22	.0045	0.00085
(Diffus.)	—	—	1.28	.0040	0.00079

$N = 50$

Rates(λ):	0.01	0.0222...	0.05	0.10	0.15
ξ(M.Ch.)	0.53	0.13	0.033	0.025	0.023
(Diffus.)	—	0.20	0.033	0.025	0.023
σ^2(M.Ch.)	0.83	0.33	0.0011	0.000081	0.000026
(Diffus.)	—	6.48	0.00099	0.000078	0.000025

The diffusion approximation and Markov chain parameters agree remarkably well when traffic is heavy (large λ), but, as might be feared, diffusion fails miserably for small λ .

6. Queue Control by Service Span and Interruption

In this section we consider queue control. The central processor now has finite service span, c , which may be smaller than the number, N , of terminals. This means that if there are $i \leq c$ jobs in service they are served as before "inside" at a rate $\mu r(i)$, with processor sharing in effect. However, if there are more than c jobs simultaneously requesting service, only c of them are served simultaneously, and at rate $\mu r(c)$, also with processor sharing discipline. The others are queued "outside", with "first-come, first-served" service discipline.

If there are more than c customers requesting service, the customers that are in service "inside" experience independent service interruptions at rate ν. When service is interrupted, each job in service is equally likely to be moved to the end of the queue; thereupon the job at the head of the "outside" queue immediately enters service. Both the imposition of the limited processor sharing, imposed by $c \leq N$, and the interruption process are intended to control queueing by adjusting the relative attention given to short and long jobs.

Markovian assumptions are made throughout, so that $X(t)$, the number of jobs requesting service at time t, is a birth and death process with transition rates given by (1.1).

(a) An Auxiliary Process.

Let R be the response time of a newly arrived tagged job. Since the tagged job may not be served until completion when it first enters service, it is necessary to introduce an auxiliary process $\{Y(t);\ t \geq 0\}$ in order to study R.

In brief summary, if there are i customers (including the tagged job) requesting service at time t and the tagged job is in service, then the state of $Y(t)$ is $(i,0)$. If there are $i > c$ customers requesting service at time t and the tagged job is not in service but is in the $j\underline{th}$ position in queue, the state of $Y(t)$ is (i,j). One can now describe the possible changes in $Y(t)$ in a time interval of length h; details will be presented elsewhere.

Let

$$m_{(i,j)}(t) = E[R \mid Y(0) = (i,j)\ ,\quad W(R) = t]\ , \tag{6.1}$$

the conditional expected response time, given that the tagged job is initially in the company of $(i-1)$ others and either it is being served inside (if $j = 0$) or it is $j\underline{th}$ in the outside queue (if $j > 0$).

Arguments similar to those of Section 2 yield differential equations for $m_{(i,j)}(t)$ which can be solved numerically. A closed-form solution is complicated and uninformative. It is possible to numerically evaluate the mean response time,

$$m(t) = E[R \mid W(R) = t] = \sum_{i,j} q_{(i,j)} m_{(i,j)}(t)\ ,$$

where $q_{(i,j)}$ is the initial distribution encountered by the tagged job.

The mean response time is not generally linear in t for this model. Note that if $c = N$ then this model is equivalent to that considered in Section 2 and hence as shown is Section 2, $m(t)$ is linear in t for that special case.

(b) An Approximation to Expected Response Time.

A useful approximation to the expected response time for a job requiring t units of work is obtained by the following argument. Assume that the service rate for the tagged job is the same throughout its processing and is equal to the rate that it experiences when it first enters the processor. Thus the tagged job requires $t^* = \frac{it}{r(i)}$ units of processing time if it enters when these are $i \leq c$ jobs (including the tagged one) requesting processing. If $i > c$, then $t^* = \frac{ct}{r(c)}$. If $i > c$, then the number of service interruptions during t is Poisson with rate $\frac{\nu}{c}$. Each time service is interrupted, the tagged job spends an expected amount of time $(i-c)\ [\nu + \mu r(c)]^{-1}$ in queue. Thus the expected time spent in queue because of service interuptions is $(i-c)(\frac{\nu}{c})t^*[\nu + \mu r(c)]^{-1}$. If $i > c$, the expected initial wait in queue until the tagged job starts service is $(i-c)[\nu + \mu r(c)]^{-1}$. The resulting approximation to the expected response time is (see (2.4))

$$A = \sum_{i=1}^{c} q_i \frac{it}{r(i)} + \sum_{i=c+1}^{N} q_i \frac{ct}{r(c)} \tag{6.2}$$

$$+ \sum_{i=c+1}^{N} q_i(i-c)[\nu + \mu r(c)]^{-1}\left[\frac{\nu t}{r(c)} + 1\right] .$$

Table 1 gives values for the expected response time, $m(t)$, and the above approximation for various values of λ, μ, ν, c, and t for $r(j) \equiv 1$ and $N = 25$ terminals. The quality of the approximation (6.2) appears to be excellent for all cases considered.

(c) Numerical Implications.

Aspects of the behavior of $m(t)$ to be noted from the table are as follows. If the amount of processing time required, t, is "small", then expected response time is minimized when c is maximized (here $c = 25$); that is, when there is maximal processor sharing and no outside queue. If t is "large", then expected response time is minimized when $c = 1$; that is, when the processor is dedicated solely to the job that is being served, and other jobs queue outside in turn. Note that increasing the rate of service interruptions by changing ν can either increase or decrease the expected response time, depending upon job time requirements.

These behavioral aspects also appear by taking derivatives of the approximate average response time A. In particular,

$$\frac{\partial}{\partial \nu} A \begin{cases} < 0 & \text{if} \quad t < \frac{1}{\mu} , \\ = 0 & \text{if} \quad t = \frac{1}{\mu} , \\ > 0 & \text{if} \quad t > \frac{1}{\mu} . \end{cases}$$

If $r(j) \equiv 1 \quad j = 1,\ldots,N$, then A is decreasing in c if $t < \frac{1}{\mu}$; A is increasing in c if $t > \frac{1}{\mu}$, and A is constant in c if $t = \frac{1}{\mu}$.

Finally, arguments similar to those in Section 3 will show that the response time is approximately normally distributed when the required work is large. Again, details will be provided in later work.

The Expected Response Time

λ	μ	ν	t	Actual	Approx.	C	1	3	5	7	9	11	13	15	17	19	21	23	25
0.1	2	0	0.1	X			2.76	2.04	1.47	1.07	0.82	0.70	0.65	0.64	0.63	0.63	0.63	0.63	0.63
					X		2.76	2.05	1.48	1.07	0.83	0.7	0.65	0.64	—	—	—	—	→
0.1	2	1	0.1	X			2.03	1.56	1.18	0.92	0.76	0.68	0.64	0.63	—	—	—	—	→
					X		2.05	1.57	1.19	.92	.76	.68	.64	.63	—	—	—	—	→
0.1	2	1	0.5	X			3.11	3.12	3.14	3.14	3.15	3.16	—	—	—	—	—	—	→
					X		3.16	—	—	—	—	—	—	—	—	—	—	—	→
1	0.5	2.0	0.4	X			10.72	10.64	10.56	10.48	10.40	10.33	10.25	10.17	10.09	10.01	9.94	9.86	9.8
					X		10.72	10.64	10.56	10.48	10.4	10.33	10.25	10.17	10.09	10.01	9.94	9.86	9.8
.01	2	1	0.1	X			0.15	0.11	—	—	—	—	—	—	—	—	—	—	→
					X		0.15	0.11	—	—	—	—	—	—	—	—	—	—	→

Table 1.

APPENDIX

To show (2.7) and (3.10) yield the same value for $E(\{R|W(\tau) = t\}$ for t large for the case $\mu_j = \mu r(j)$.

First:

$$q(j) = c\pi_j \mu_j$$

$$= c \, \frac{\lambda_0 \times \ldots \times \lambda_{j-1}}{\mu_1 \times \ldots \times \mu_{j-1}} \, .$$

Hence,

$$m \equiv \sum_{j=1}^{\infty} \frac{j}{r(j)} q_j$$

$$= c \sum_{j=1}^{\infty} \frac{j}{r(j)} \frac{\lambda_0 \times \ldots \times \lambda_{j-1}}{\mu_1 \times \ldots \times \mu_{j-1}}$$

Since,

$$\mu_j = \mu r(j)$$

$$m = c\mu \sum_{j=1}^{\infty} j \, \frac{(\lambda_0 \times \ldots \times \lambda_{j-1})}{\mu_1 \times \ldots \times \mu_j}$$

$$= \frac{c\mu}{\prod_{k=1}^{\infty} \mu_k} \sum_{j=1}^{\infty} j(\lambda_0 \times \ldots \times \lambda_{j-1}) \left(\prod_{i=j+1}^{\infty} \mu_i \right)$$

Further,

$$c = \frac{\prod_{k=1}^{\infty} \mu_k}{\sum_{j=1} \lambda_0 \times \ldots \times \lambda_{j-1} \prod_{k=j}^{\infty} \mu_k} \, .$$

Hence,

$$m = \mu \, \frac{\sum_{j=1}^{\infty} j(\lambda_0 \times \ldots \times \lambda_{j-1}) \left(\prod_{k=j+1}^{\infty} \mu_k \right)}{\sum_{j=1}^{\infty} (\lambda_0 \times \ldots \times \lambda_{j-1}) \left(\prod_{k=j}^{\infty} \mu_k \right)} = \frac{\sum_{j=1}^{\infty} j \left(\prod_{i=1}^{j-1} \lambda_i \right) \left(\prod_{k=j+1}^{\infty} \mu_k \right)}{\sum_{j=1}^{\infty} \left(\prod_{i=1}^{j-1} \lambda_i \right) \left(\prod_{k=j}^{\infty} \mu_k \right)}$$

where $\lambda_1 \times \lambda_0 = 1$ by convention.

__Second:__

$$\pi'(j) = \frac{\lambda_1 \times \dots \times \lambda_j}{\mu_2 \times \dots \times \mu_{j+1}} (j+1)\pi'(0) \qquad j = 0,1,\dots \qquad [\lambda_1 \times \lambda_0 \equiv 1]$$

where

$$\pi'(0) = \frac{\prod_{k=2}^{\infty} \mu_k}{\sum_{j=1}^{\infty} j(\lambda_1 \times \dots \times \lambda_{j-1}) \prod_{k=j+1}^{\infty} \mu_k}$$

$$\xi = \sum_{j=0}^{\infty} \frac{r(j+1)}{j+1} \pi'(j) = \sum_{j=1}^{\infty} \frac{r(j)}{j} \pi'(j-1)$$

$$= \sum_{j=1}^{\infty} r(j) \frac{\lambda_1 \times \dots \times \lambda_{j-1}}{\mu_2 \times \dots \times \mu_j} \pi'(0)$$

$$= \frac{1}{\mu} \sum_{j=1}^{\infty} \frac{\lambda_1 \times \dots \times \lambda_{j-1}}{\mu_2 \times \dots \times \mu_{j-1}} \pi'(0)$$

since

$$\mu_{j+1} = \mu r(j+1) \text{ . Thus}$$

$$\xi = \frac{\pi'(0)}{\left[\prod_{k=2}^{\infty} \mu_k\right]} \frac{1}{\mu} \sum_{j=1}^{\infty} \lambda_1 \times \dots \times \lambda_{j-1} \left[\prod_{k=j}^{\infty} \mu_k\right]$$

$$= \frac{\sum_{j=1}^{\infty} (\lambda_1 \times \dots \times \lambda_{j-1}) \left[\prod_{k=j}^{\infty} \mu_k\right]}{\mu\left[\sum_{j=1}^{\infty} j(\lambda_1 \times \dots \times \lambda_{j-1}) \prod_{k=j+1}^{\infty} \mu_k\right]}$$

A comparison of ξ with m shows that

$$m = \frac{1}{\xi} .$$

REFERENCES

Burman, D. (1979), "An analytic approach to diffusion approximations in queueing". Unpublished Doctoral Dissertation, New York University.

Cinlar, E. (1975), Introduction to Stochastic Processes, Prentice-Hall, Englewood Cliffs, N.J.

Coffman, E. G., Muntz, R. R. and Trotter, H. (1970), "Waiting time distribution for processor-sharing systems," J. Assn. for Comp. Mach., 17, pp. 123-130.

Cox, D. R. (1962), Renewal Theory, Methuen Monograph.

Gaver, D. P., and Lehoczky, J. P. (1976), "Gaussian approximation to service problems: a communications system example," J. Appl. Prob., 13, pp. 768-780.

Gaver, D., Jacobs, P. and Latouche, G. (1981). "Finite birth and death models in randomly changing environments". To appear J. Appl. Prob.

Iglehart, D. L. (1965), "Limiting diffusion approximations for the many-server queue and the repairman problem," J. Appl. Prob., 2, pp.

Karlin, S., and Taylor, H. M. (1975), A First Course in Stochastic Processes, (Second Edition). Academic Press, New York.

Keilson, J. (1979), Markov Chain Models - Rarity and Exponentiality, Springer-Verlag, New York.

Kelly, F. P. (1979), Reversibility and Stochastic Networks, John Wiley and Sons, New York.

Kleinrock, L. (1976), Queueing Systems, Vol. II, Wiley-Interscience.

Lavenberg, S. S. and Reiser, M. (1980), "Stationary state probabilities at arrival instants for closed queueing networks with multiple types of customers," J. Appl. Prob., 17, pp. 1048-1061.

Mitra, D. (1981), "Waiting time distributions from closed queueing network models of shared processor systems," Bell Laboratories Report.

Cohen, J. W. (1979), "The multiple phase service network with generalized processor sharing. Acta Informatica 12 245-284.

DISTRIBUTION LIST

	NO. OF COPIES
Library, Code 0142 Naval Postgraduate School Monterey, CA 93943	2
Director of Research Administration Code 012A Naval Postgraduate School Monterey, CA 93943	1
Library, Code 55 Naval Postgraduate School Monterey, CA 93943	1
Defense Technical Information Center Cameron Station Alexandria, VA 22314	2
Prof. M. L. Abdel-Hameed Department of Mathematics University of North Carolina Charlotte, NC 28223	1
Dr. G. P. Alldredge Department of Physics The University of Missouri Columbia, MO 65211	1
Prof. F. J. Anscombe Department of Statistics Yale University, Box 2179 New Haven, CT 06520	1
Dr. Barbara Bailar Associate Director Statistical Standards Bureau of Census Washington, DC 20024	1
Mr. C. M. Bennett Code 741 Naval Coastal Systems Laboratory Panama City, FL 32401	1
Dr. Derrill J. Bordelon Code 21 Naval Underwater Systems Center Newport, RI 02840	1

DISTRIBUTION LIST

	NO. OF COPIES
Dr. David Brillinger Statistics Department University of California Berkeley, CA 94720	1
Dr. R. W. Butterworth Systems Exploration 1340 Munras Avenue Monterey, CA 93940	1
Dr. D. R. Cox Department of Mathematics Imperial College London SW7 ENGLAND	1
Dr. D. F. Daley Statistics Dept. (IAS) Australian National University Canberra A.C.T. 2606 AUSTRALIA	1
Mr. DeSavage Naval Surface Weapons Center Silver Springs, MD 20910	1
Professor C. Derman Dept. of Civil Eng. & Mech. Engineering Columbia University New York, NY 10027	1
Dr. Guy Fayolle I.N.R.I.A. Dom de Voluceau-Rocquencourt 78150 Le Chesnay Cedex FRANCE	1
Dr. M. J. Fischer Defense Communications Agency 1860 Wiehle Avenue Reston, VA 22070	1
Professor George S. Fishman Cur. in OR & Systems Analysis University of North Carolina Chapel Hill, NC 20742	1
Dr. R. Gnanadesikan Bell Telephone Lab Murray Hill, NJ 07733	

DISTRIBUTION LIST

	NO. OF COPIES
Professor Bernard Harris Department of Statistics University of Wisconsin 610 Walnut Street Madison, WI 53706	1
Dr. Gerhard Heiche Naval Air Systems Command (NAIR 03) Jefferson Plaza, No. 1 Arlington, VA 20360	1
Professor L. H. Herbach Department of Mathematics Polytechnic Institute of NY Brooklyn, NY 11201	1
Professor W. M. Hinich University of Texas Austin, TX 78712	1
P. Heidelberger IBM Research Laboratory Yorktown Heights New York, NY 10598	1
W. D. Hibler, III Geophysical Fluid Dynamics Princeton University Princeton, NJ 08540	1
Professor D. L. Iglehart Department of Operations Research Stanford University Stanford, CA 94350	1
Dr. D. Vere Jones Department of Mathematics Victoria University of Wellington P. O. Box 196 Wellington NEW ZELAND	1
Professor J. B. Kadane Department of Statistics Carnegie-Mellon Pittsburgh, PA 15213	1

DISTRIBUTION LIST

NO. OF COPIES

Professor Guy Latouche 5
University Libre Bruxelles
C.P. 212
Blvd De Triomphe
B-1050 Bruxelles
BELGIUM

Dr. Richard Lau 1
Office of Naval Research
Branch Office
1030 East Green Street
Pasadena, CA 91101

A. J. Laurance 1
Department of Mathematics Statistics
University of Birmingham
P. O. Box 363
Birmingham B15 2TT
ENGLAND

Dr. John Copas 1
Department of Mathematics Statistics
University of Birmingham
P. O. Box 363
Birmingham B15 2TT
ENGLAND

Professor M. Leadbetter 1
Department of Statistics
University of North Carolina
Chapel Hill, NC 27514

Mr. Dan Leonard 1
Cod 8105
Naval Pcean Systems Center
San Diego, CA 92152

M. Lepparanta 1
Winter Navigation Res. Bd.
Helsinki
FINLAND

J. Lehoczky 1
Department of Statistics
Carnegie-Mellon University
Pittsburgh, PA 15213

Library 1
Naval Ocean Systems Center
San Diego, CA 92152

DISTRIBUTION LIST

NO. OF COPIES

Dr. J. Maar (R51) 1
National Security Agency
Fort Meade, MD 20755

Bob Marcello 1
Canada Marine Engineering
Calgary
CANADA

Dr. M. McPhee 1
Chair of Arctic Marine Science
Oceanography Department
Naval Postgraduate School
Monterey, CA 93943

Dr. M. Mazumdar 1
Dept. of Industrial Engineering
University of Pittsburgh
Oakland
Pittsburgh, PA 15235

Professor Rupert G. Miller, Jr. 1
Statistics Department
Sequoia Hall
Stanford University
Stanford, CA 94305

National Science Foundation 1
Mathematical Sciences Section
1800 G Street, NW
Washington, DC 20550

Naval Research Laboratory 1
Technical Information Section
Washington, DC 20375

Professor Gordon Newell 1
Department of Civil Engineering
University of California
Berkeley, CA 94720

Dr. David Oakes 1
TUO Centenary Inst. of Occ. Health
London School of Hygiene/Tropical Med.
Keppel St. (Gower St.)
London WO1 E7H1,
ENGLAND

DISTRIBUTION LIST

NO. OF COPIES

Dr. Alan F. Petty 1
Code 7930
Navy Research Laboratory
Washington, DC 20375

E. M. Reimnitz 1
Pacific-Arctic Branch-Marine Geology
U.S. Geological Survey
345 Middlefield Rd., (MS99)
Menlo Park, CA 94025

Professor M. Rosenblatt 1
Department of Mathematics
University of California - San Diego
La Jolla, CA 92093

Professor I. R. Savage 1
Department of Statistics
Yale University
New Haven, CT 06520

Professor W. R. Schucany 1
Department of Statistics
Southern Methodist University
Dallas, TX 75222

Professor D. C. Siegmund 1
Department of Statistics
Sequoia Hall
Stanford University
Stanford, CA 94305

Professor H. Solomon 1
Department of Statistics
Sequoia Hall
Stanford University
Stanford, CA 94305

Dr. Ed Wegman 1
Statistics & Probability Program
Code 411(SP)
Office of Naval Research
Arlington, VA 22217

Dr. Douglas de Priest 1
Statistics & Probability Program
Code 411(SP)
Office of Naval Research
Arlington, VA 22217

DISTRIBUTION LIST

NO. OF COPIES

Dr. Marvin Moss 1
Statistics & Probability Program
Code 411(SP)
Office of Naval Research
Arlington, VA 22217

Technical Library 1
Naval Ordnance Station
Indian Head, MD 20640

Professor J. R. Thompson 1
Dept. of Mathematical Science
Rice University
Houston, TX 77001

Professor J. W. Tukey 1
Statistics Department
Princeton University
Princeton, NJ 08540

P. Wadhams 1
Scott Polar Research
Cambridge University
Cambridge CB2 1ER
ENGLAND

Daniel H. Wagner 1
Station Square One
Paoli, PA 19301

Dr. W. Weeks 1
U.S. Army CR REL
72 Lyme Road
Hanover, NH 03755

P. Welch 1
IBM Research Laboratory
Yorktown Heights, NY 10598

Pat Welsh 1
Head, Polar Oceanography Branch
Code 332
Naval Ocean Research & Dev. Activity
NSTL Station
Mississippi 39529

Dr. Roy Welsch 1
Sloan School
M.I.T.
Cambridge, MA 02139

DISTRIBUTION LIST

	NO. OF COPIES
Dr. Morris DeGroot Statistics Department Carnegie-Mellon University Pittsburgh, PA 15235	1
Professor R. Renard Chairman Meteorology Department Naval Postgraduate School Monterey, CA 93943	1
Dr. A. Weinstein Commanding Officer Naval Environmental Prediction Research Facility Monterey, CA 93943	1
Paul Lowe Naval Environmental Prediction Research Facility Monterey, CA 93943	1
Wayne Sweet Naval Environmental Prediction Research Facility Monterey, CA 93943	1
Dr. Colin Mallows Bell Telephone Laboratories Murray Hill, NJ 07974	1
Dr. D. Pregibon Bell Telephone Laboratories Murray Hill, NJ 07974	1
Dr. Jon Kettenring Bell Telephone Laboratories Murray Hill, NJ 07974	1
Professor D. P. Gaver Code 55Gv Naval Postgraduate School Monterey, CA 93943	15
Associate Professor P. A. Jacobs Code 55Jc Naval Postgraduate School Monterey, CA 93943	10